Recall 2.1: Visualize variation in numerical data

- Dot plot
- Histogram
- Density plot

Recall 2.2: Features of numerical distribution

- Shape (symmetry, peaks, outliers)
- Center (typical value)
- Spread (diversity, variability)

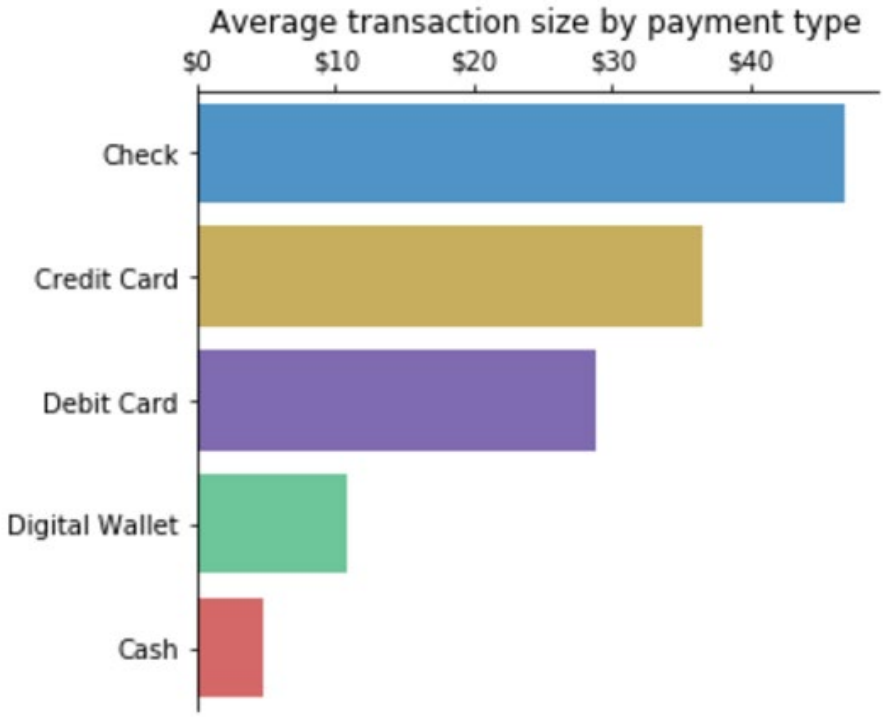
Section 2.3

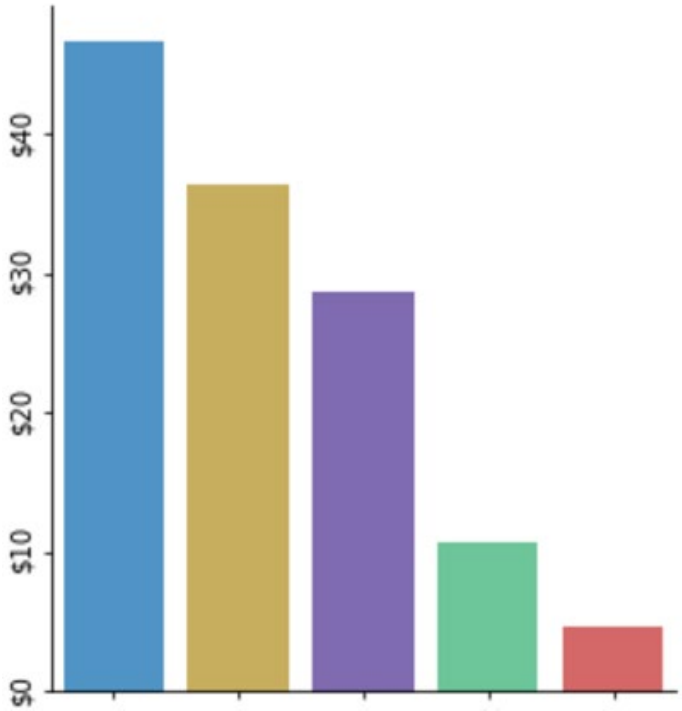
Visualizing Variation in Categorical Data

- Bar chart (bar graph, bar plot)
- Pie chart (pie graph)

2.3: Visualizing Variation in Categorical Data

- Bar chart (bar graph, bar plot)





Bar Chart

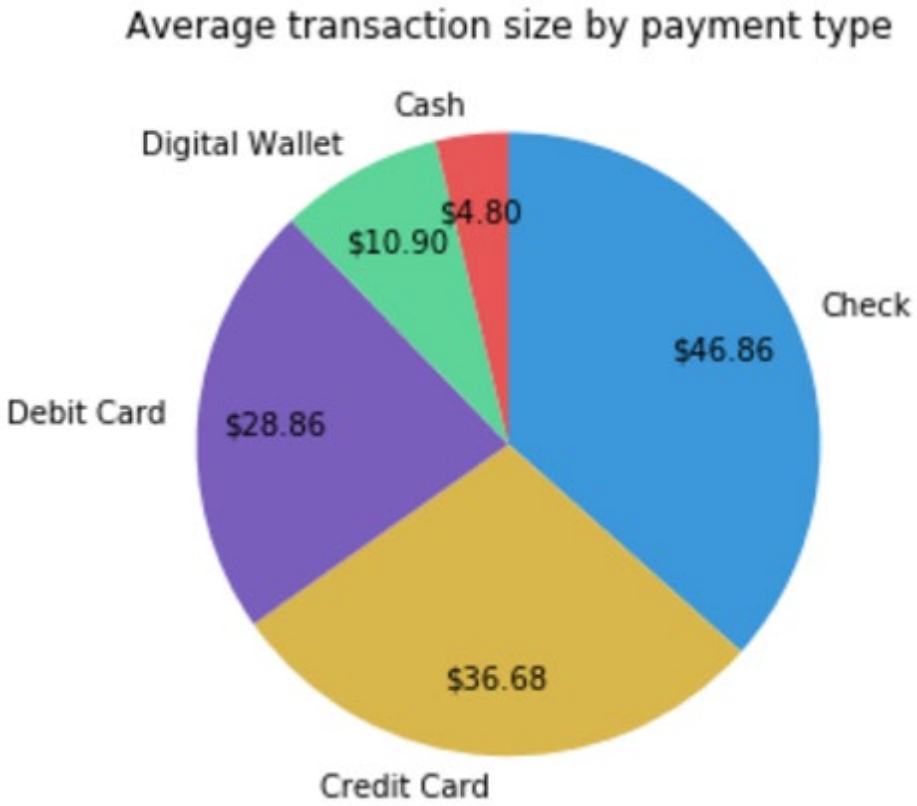
We treat categorical data similar to numerical data.

Bar charts and histograms look similar, but they have some key differences. In a bar chart,

- The order of bars doesn't matter because the categories have no natural order.
- The widths of the bars have no meaning.
- The gaps between the bars have no meaning.

2.3: Visualizing Variation in Categorical Data

- Pie chart (pie graph, pie plot)



Pie Chart

A **pie chart** looks like a pie.

- The pie is sliced into several pieces, and each piece represents a category of the variable.
- The area of the piece is proportional to the relative frequency of that category.

Section 2.4
Summarizing Categorical Distributions

Recall:

> *To describe a numerical distribution, we record shape, center, and spread.*

Categorical data has no inherent order. These measures do not make sense for categorical data.

Two main components of a categorical distribution:

- *Mode: typical (most frequent) category*
- *Variability: diversity in categories*

Mode

Mode is the category that occurs the most frequently.

Differences in the mode for categorical and numerical data:

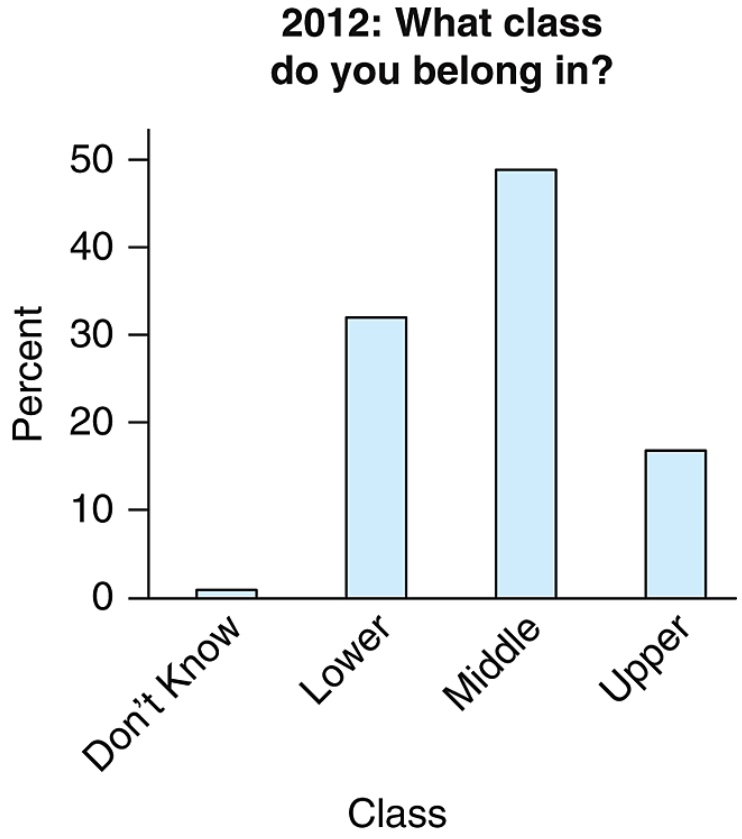
- Mounds should be prominent but do not need to be of the same height.
- Modes must be roughly the same height.

We use the same wording as before:

- Unimodal: one distinct mode
- Bimodal: two modes with same (or very close) frequency
- Multimodal: more than two modes with (or close) frequency

Mode: Example

In 2012, the Pew survey asked a new group of 2508 Americans which economic class they identified with.

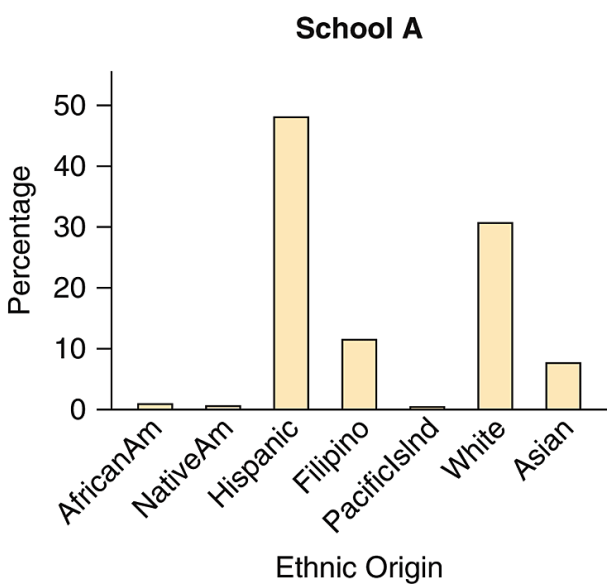


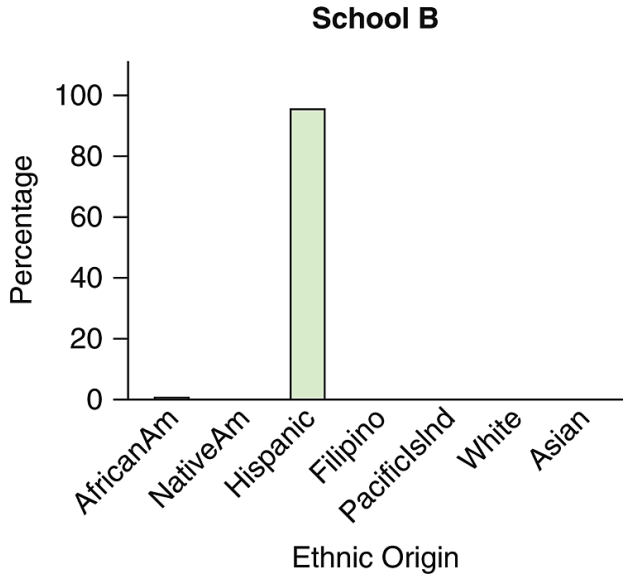
Variability

- Variability is the diversity in the data values rather than in frequencies.
- Numerical data: variability is determined by the range (horizontal spread) of the data.
- Categorical data: variability is determined by the number of different categories.

Variability: Example

The bar charts below show the ethnic composition of two schools in the Los Angeles City School System.



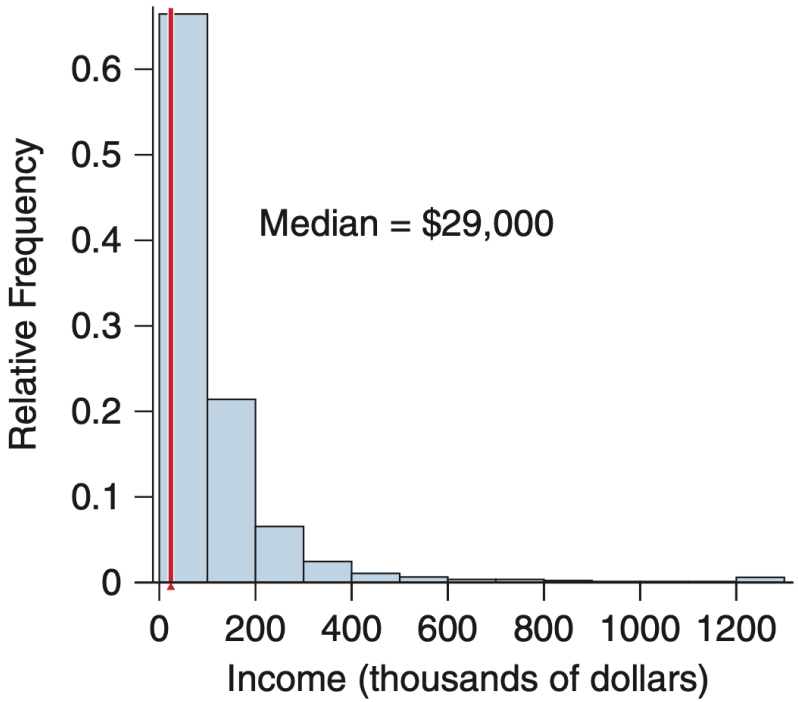


School A has the greater variability in ethnicity.

Section 3.3:
Summaries for Skewed Distributions

- Measure for center: Median
- Measure of variability: Interquartile Range (IQR)

Median



The histogram shows the distribution of incomes for a sample of New York State residents. About half the residents have incomes above $29,000 and about half have incomes below $29,000.

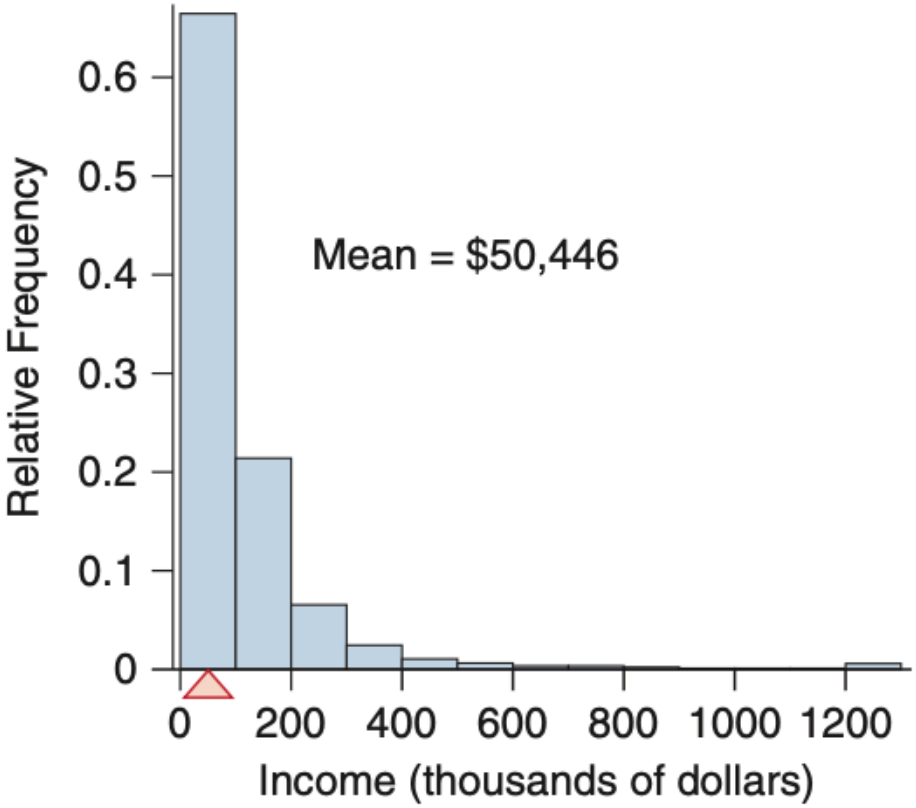
Median

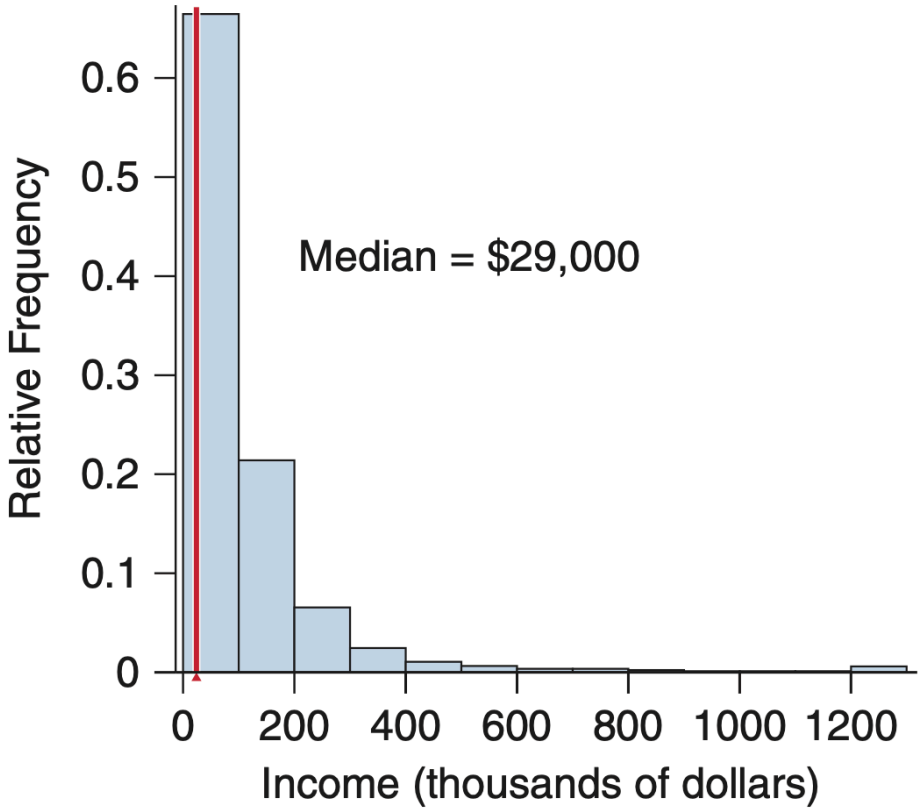
Median: the middle number

- Arrange all numbers in order.
- Find the middle number. If there are two numbers in the middle, average them.

In a skewed distribution, the median is better than mean as a typical value because median is less sentitive to outliers.

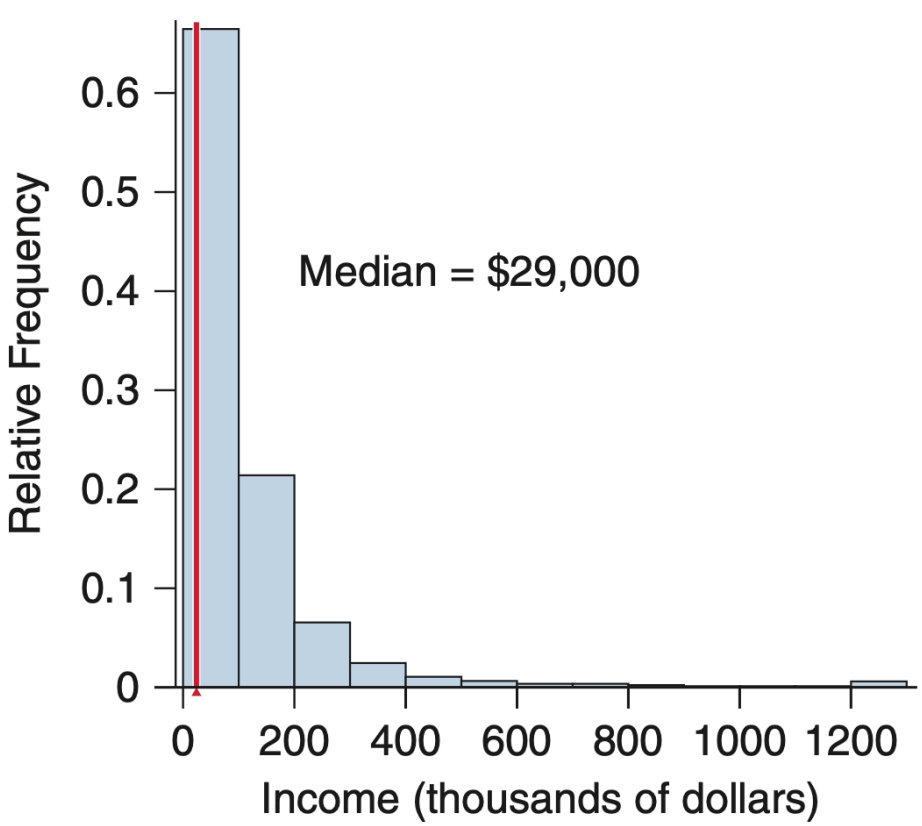
Typical Value: The Mean vs. the Median

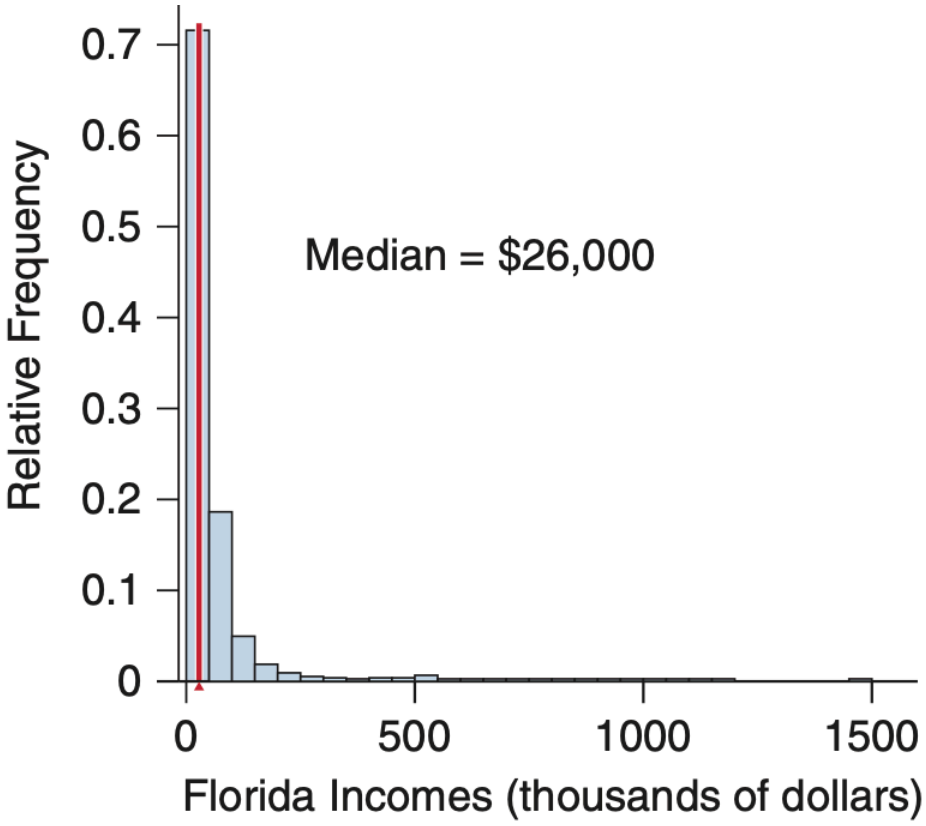




Which is a better measure of the "typical" income of a New York State resident: the mean or the median?

The Median: Comparing Groups





New York State resident: median is $29,000

Florida resident: median is $26,000

The Median: Example

The prices of a gallon of regular gas at 12 gas stations are the following:

$2.69, $2.69, $2.79, $2.85, $2.85, $2.89, $2.89, $2.89, $2.86, $2.89, $2.89, $2.89

Find the median price for a gallon of gas and interpret the value.

The Median: Example

Arrange the data values:

2.69, 2.69, 2.79, 2.85, 2.85, 2.86, 2.89, 2.89, 2.89, 2.86, 2.89, 2.89, 2.89

The median is (2.86+2.89)/2=2.875.

This the typical price of a gallon of gas at these 12 gas stations.

Measuring the Variability

The *standard deviation* (STD) is a measure of spread using the distance from the mean.

The *interquartile range* (IQR) is a measure of spread related to the median.

Range

Range: The difference between the largest and smallest values

Example:

A group of eight children have the following heights:

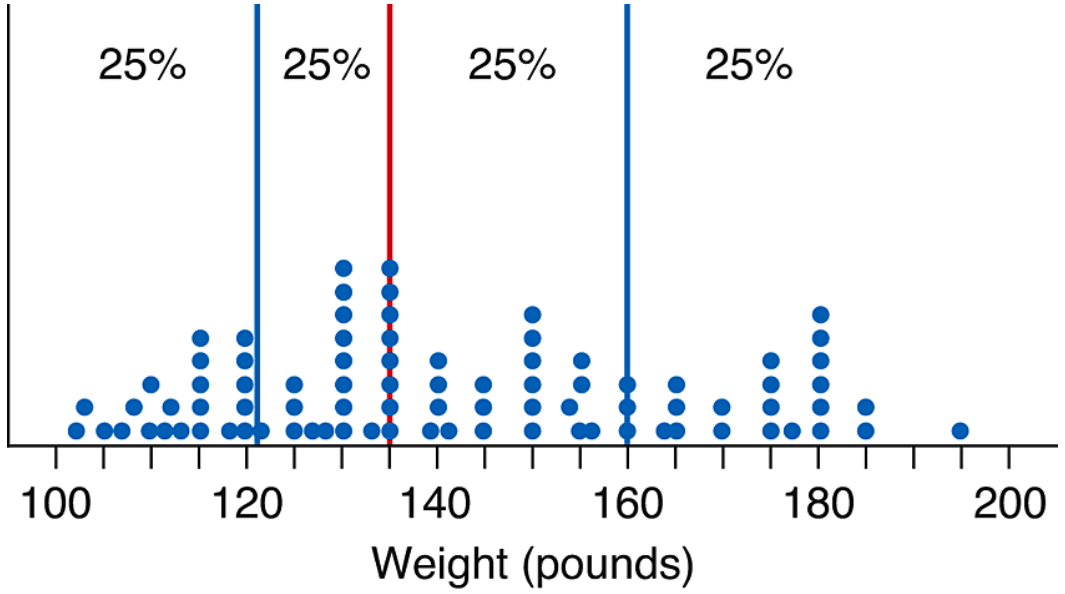
48.0, 48.0, 53.0, 53.5, 54.0, 60.0, 62.0, and 71.0

The range in the children's heights is 71.0 − 48.0 = 23.0 inches.

Quartiles

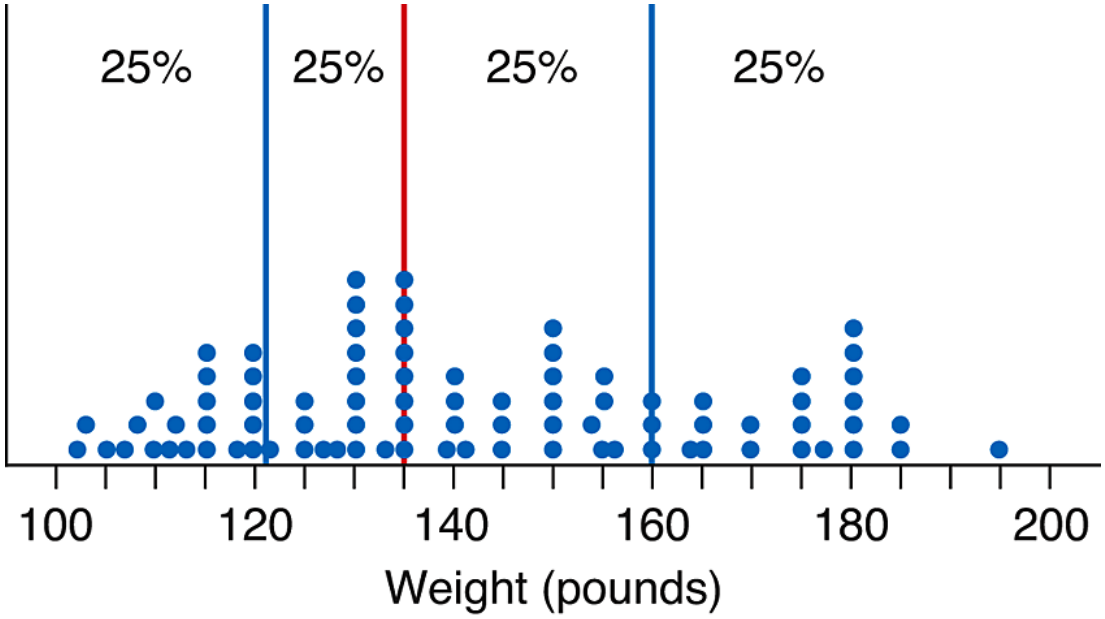
Quartiles divide the distribution into fourths. Each quartile contains 25% of the data.

Example: The dotplot shows the distribution of weights for a class of introductory statistics students.



Interquartile Range: IQR

IQR = the range of the middle 50% of the data



$$\text{IQR} = 160 - 121 = 39 \text{ pounds}$$

(distance between the first and third “slice”)